

A Apprentissage artificiel

Antoine Cornuéjols - Laurent Miclet

Avec la participation d'Yves Kodratoff

Concepts et algorithmes

Préface de Tom Mitchell

EYROLLES

Table des matières

Avant-propos	v
Quelques applications de l'apprentissage artificiel	vi
Quelques définitions de base	vi
Deux champs industriels de l'apprentissage artificiels : la reconnaissance des formes et la fouille de données	vii
Les caractéristiques de l'apprentissage artificiel	viii
Trois exemples d'apprentissage	x
Organisation et plan de l'ouvrage	xiii
Guide de lecture	xiv
La situation de l'apprentissage dans l'intelligence artificielle	xv
Les applications industrielles de l'apprentissage artificiel à l'extraction de connaissances des données	xv
Notations	xxvii
I Les fondements de l'apprentissage	1
1 De l'apprentissage naturel à l'apprentissage artificiel	3
1.1 L'apprentissage artificiel	3
1.2 Deux exemples : apprendre à jouer, apprendre à lire	5
1.2.1 Apprendre à jouer	5
1.2.2 Apprendre à reconnaître des caractères manuscrits	7
1.3 Deux approches : la cybernétique et les sciences cognitives	9
1.3.1 La cybernétique	10
1.3.2 Le pari du cognitivisme	11
1.4 Les concepts de base de l'apprentissage	13
1.4.1 Un scénario de base pour l'induction	13
1.4.2 Quelques notions clés	14
1.4.3 L'induction vue comme une estimation de fonction	18
1.5 L'induction comme un jeu entre espaces	20
1.5.1 L'apprentissage est impossible...	22
1.5.2 ... sans limiter l'espace des hypothèses	23
1.5.3 L'exploration de l'espace des hypothèses	26
1.6 Retour sur l'organisation de l'ouvrage	27

2	Première approche théorique de l'induction	35
2.1	Deux exemples d'induction	37
2.1.1	Le système ARCH	37
2.1.2	Le perceptron	39
2.2	Approche de l'induction	42
2.2.1	Le compromis biais-variance	42
2.2.2	Comment définir formellement le problème de l'induction?	45
2.2.3	Quel principe inductif adopter? Une introduction	46
2.2.4	Comment analyser l'apprentissage?	48
2.3	Analyse dans le pire cas : l'apprentissage <i>PAC</i>	49
2.3.1	Étude des conditions de validité de l' <i>ERM</i>	50
2.3.2	Le cas de la discrimination : l'analyse <i>PAC</i>	53
2.4	Analyse dans un cas moyen : l'analyse bayésienne	57
2.4.1	Nature de l'analyse bayésienne	58
2.4.2	Le risque bayésien et la décision optimale	59
2.4.3	Cas particuliers de la décision bayésienne	60
2.4.4	Panorama des méthodes inductives dans le cadre bayésien	63
2.4.5	Et si l'espace des hypothèses ne contient pas la fonction cible?	63
2.4.6	En résumé : la procédure inductive bayésienne	64
2.5	Discussion : Quels types d'analyses et de principes inductifs?	64
2.6	Les grands principes inductifs avec régulation des hypothèses	65
2.6.1	L'idée générale : le réglage de la classe d'hypothèses	66
2.6.2	La sélection de modèles	67
2.7	Discussion et perspectives	68
2.8	Notes historiques et bibliographiques	69
3	L'environnement méthodologique de l'apprentissage	73
3.1	L'espace des données d'apprentissage	76
3.1.1	La représentation des objets de l'apprentissage	76
3.1.2	Le prétraitement des données	80
3.2	L'espace des hypothèses d'apprentissage	86
3.2.1	Le problème général de la représentation des connaissances	86
3.2.2	La classification	87
3.2.3	La régression	90
3.2.4	Les distributions de probabilités	90
3.2.5	Les arbres de décision	90
3.2.6	Les hiérarchies de concepts	91
3.2.7	Les réseaux bayésiens et les modèles graphiques	92
3.2.8	Les chaînes de Markov et les modèles de Markov cachés	93
3.2.9	Les grammaires	93
3.2.10	Les formalismes logiques	94
3.3	La recherche dans l'espace des hypothèses	96
3.3.1	Caractérisation de l'espace de recherche	96
3.3.2	Caractérisation des fonctions de coût	96
3.3.3	Les méthodes d'optimisation	97
3.4	L'évaluation de l'apprentissage	102
3.4.1	L'évaluation <i>a priori</i> : critères théoriques	103
3.4.2	L'évaluation empirique <i>a posteriori</i> : généralités	104

3.4.3	Risque empirique et risque réel	105
3.4.4	La sélection de modèle en pratique	106
3.4.5	L'estimation du risque réel d'une hypothèse	111
3.4.6	Le réglage des algorithmes par un ensemble de validation	115
3.4.7	D'autres critères d'appréciation	117
3.5	La comparaison des méthodes d'apprentissage	118
3.5.1	La comparaison de deux hypothèses produites par un même algorithme sur deux échantillons de test différents.	119
3.5.2	La comparaison de deux algorithmes sur des ensembles de test différents .	119
3.5.3	La comparaison de deux algorithmes sur le même ensemble de test	120
II Apprentissage par exploration		123
4	Induction et relation d'ordre : l'espace des versions	125
4.1	Les concepts de base	128
4.1.1	La description des attributs, la description des concepts	128
4.1.2	Les sélecteurs	128
4.1.3	La relation de généralité entre les hypothèses	129
4.1.4	La relation entre un objet et un concept	131
4.2	La structuration de l'espace des hypothèses	132
4.2.1	Preliminaires	132
4.2.2	Un exemple: les paires de rectangles	133
4.2.3	Un ordre partiel sur l'espace des hypothèses	134
4.2.4	Quelques opérateurs de spécialisation et de généralisation	136
4.2.5	Quelques propriétés utiles d'un espace structuré par une relation d'ordre partiel	137
4.3	La construction de l'espace des versions	139
4.3.1	Illustration: retour sur l'exemple des rectangles	139
4.3.2	L'algorithme d'élimination des candidats	140
4.3.3	Deux exemples	140
4.3.4	Un exemple d'application: le système LEX	148
4.4	Analyse de l'algorithme d'élimination de candidats	149
4.4.1	Complexité au pire	149
4.4.2	Le point de vue de l'apprentissage <i>PAC</i>	150
4.5	La représentation des connaissances par un treillis de Galois	151
4.5.1	La construction de la structure	151
4.5.2	L'utilisation pour l'apprentissage	153
5	La programmation logique inductive	157
5.1	La programmation logique inductive: le cadre général	160
5.1.1	Complexité de l'induction et expressivité du langage d'hypothèses	160
5.1.2	La relation de couverture en logique du premier ordre	161
5.1.3	La subsomption en logique du premier ordre	163
5.1.4	Un résumé des relations de subsomption possibles	165
5.2	La logique des prédicats et les programmes logiques: terminologie	166
5.2.1	La syntaxe de la logique des prédicats	166

5.2.2	Système de preuve pour les langages de clauses	168
5.3	La structuration de l'espace des hypothèses en logique des prédicats	170
5.3.1	Le calcul de la <i>lgg</i> pour la θ -subsomption	170
5.3.2	Le calcul de <i>rlgg</i> pour la θ -subsomption relative	172
5.3.3	Le calcul de <i>lgg</i> pour la résolution inverse	174
5.4	L'exploration de l'espace des hypothèses	177
5.4.1	Le squelette des algorithmes de PLI	178
5.4.2	Les biais de recherche dans l'espace d'hypothèses	180
5.5	Deux exemples de systèmes de PLI	182
5.5.1	Un système empirique descendant : FOIL	182
5.5.2	Un système empirique ascendant : PROGOL	184
5.6	Les domaines d'application de la PLI	186
5.7	Les chantiers de la PLI	188
5.7.1	Les problèmes à résoudre	188
6	Reformulation et transfert de connaissances	193
6.1	L'apprentissage en présence de théorie	194
6.2	L'apprentissage par examen de preuve (<i>EBL</i>)	194
6.2.1	Le principe de l' <i>EBL</i>	194
6.2.2	Une illustration de l'apprentissage <i>EBL</i>	195
6.2.3	Discussion sur l'apprentissage de concept à partir d'explications	198
6.2.4	L'apprentissage de connaissances de contrôle à partir d'explications	199
6.2.5	Bilan sur l'apprentissage à partir d'explications	201
6.3	Abstraction et reformulation des connaissances	201
6.4	Changement de repère et raisonnement par analogie	203
6.5	Bilan	205
7	L'inférence grammaticale	207
7.1	Définitions et notations	212
7.1.1	Langages, grammaires, automates et partitions	212
7.1.2	Échantillons d'un langage et automates associés	218
7.2	Les protocoles de l'inférence : quelques résultats théoriques	220
7.2.1	La spécification d'un problème d'inférence grammaticale	220
7.2.2	L'identification à la limite d'une grammaire	221
7.2.3	Deux propriétés de l'identification à la limite.	222
7.2.4	Autres protocoles pour l'inférence de grammaires.	222
7.2.5	L'inférence grammaticale et l'apprentissage <i>PAC</i>	223
7.2.6	Résultats <i>PAC</i> pour les langages réguliers	224
7.2.7	Apprentissage <i>PACS</i> : <i>PAC</i> Simple	225
7.2.8	Apprentissage <i>PAC</i> avec distributions bienveillantes	226
7.3	L'espace de recherche de l'inférence régulière	226
7.3.1	Le point de la situation	226
7.3.2	Deux propriétés fondamentales	227
7.3.3	La taille de l'espace de recherche	228
7.4	L'inférence régulière sans échantillon négatif	228
7.4.1	Une méthode caractérisable : l'inférence de langages k-réversibles	228
7.4.2	Une méthode heuristique : l'algorithme ECGI	230
7.5	L'inférence régulière sous contrôle d'un échantillon négatif	233

7.5.1	L'ensemble frontière	232
7.5.2	Le lien avec l'espace des versions	233
7.5.3	Les algorithmes RIG et BRIG	233
7.5.4	L'algorithme RPNI	234
7.5.5	Variantes et extensions	236
7.6	L'inférence de grammaires algébriques.	237
7.6.1	Présentation	237
7.6.2	L'apprentissage à partir d'échantillons structurés.	237
7.6.3	Les méthodes par exploration.	239
7.6.4	Une méthode avec oracle.	239
7.6.5	L'inférence de grammaires linéaires équilibrées	240
7.7	Quelques extensions	240
7.7.1	Les grammaires stochastiques	240
7.7.2	Le point de vue connexionniste	242
8	Apprentissage par évolution simulée	245
8.1	Trois espaces au lieu de deux	247
8.2	Un modèle formel simplifié de l'évolution	249
8.2.1	Le jeu entre \mathcal{H} et \mathcal{G}	249
8.2.2	L'apprentissage comme processus d'évolution d'une population	249
8.3	Les algorithmes génétiques	250
8.3.1	La représentation dans l'espace génotypique	251
8.3.2	L'algorithme générique	251
8.3.3	L'initialisation de la population	252
8.3.4	Les opérateurs	252
8.3.5	La sélection	255
8.3.6	Le théorème des schémas : une explication de la puissance des AG ?	258
8.4	Les stratégies d'évolution	262
8.5	La programmation génétique	262
8.5.1	La représentation des programmes	265
8.5.2	Les opérateurs génétiques sur les programmes	267
8.5.3	Évaluation et sélection	267
8.5.4	Le fonctionnement	267
8.5.5	Illustrations	268
8.5.6	Une analyse de la programmation génétique	272
8.6	La coévolution	273
8.6.1	Un exemple d'écologie : les systèmes de classeurs	274
III	Apprentissage par optimisation	277
9	L'apprentissage de surfaces séparatrices linéaires	279
9.1	Généralités.	281
9.1.1	Hyperplans séparateurs et discriminants dans un problème à deux classes.	281
9.1.2	Un peu de géométrie dans \mathbb{R}^d	282
9.2	L'apprentissage d'un hyperplan pour discriminer deux classes	283
9.2.1	Une solution globale	283
9.2.2	Une méthode itérative : l'algorithme de Ho et Kashyap.	286

9.2.3	Un autre calcul : l'algorithme du perceptron	287
9.2.4	L'hyperplan discriminant de Fisher	290
9.2.5	Surfaces séparatrices non linéaires	292
9.2.6	Et pour plus de deux classes?	292
9.3	Les séparateurs à vastes marges (SVM)	293
9.3.1	La recherche des séparateurs linéaires à vastes marges	294
9.3.2	Quelle justification pour les SVM?	302
9.3.3	La régression par fonctions noyau et exemples critiques	307
9.3.4	Conclusions sur les SVM	309
10	L'apprentissage de réseaux connexionnistes	311
10.1	Les différents éléments d'un réseau connexionniste	313
10.2	L'architecture multicouche	315
10.2.1	La transmission de l'information dans un réseau multicouche	315
10.2.2	Un exemple	316
10.2.3	Un autre exemple : le problème « XOR »	318
10.2.4	Le protocole d'apprentissage	319
10.2.5	Le codage des exemples d'apprentissage	320
10.3	L'algorithme d'apprentissage	320
10.3.1	Retour sur le perceptron	321
10.3.2	L'apprentissage par rétropropagation du gradient de l'erreur	325
10.3.3	L'organisation des calculs	326
10.3.4	Retour sur l'exemple	326
10.3.5	Une variante	328
10.3.6	Quand arrêter l'apprentissage?	329
10.3.7	Le problème des minima locaux	329
10.4	Quelques résultats théoriques sur les réseaux connexionnistes	329
10.4.1	Pouvoir d'expression	329
10.4.2	Complexité	330
10.4.3	Réseaux connexionnistes et apprentissage <i>PAC</i>	330
10.5	Comment choisir l'architecture d'un réseau?	331
11	Apprentissage par combinaison de décisions	333
11.1	Les arbres de décision	335
11.1.1	Principe	335
11.1.2	La construction récursive d'un arbre de décision	336
11.1.3	Comment élaguer un arbre trop précis	343
11.1.4	Un exemple : les iris de Fisher	347
11.1.5	Traduction des arbres de décision en logique des propositions	349
11.2	Les arbres de régression	352
11.2.1	Le principe	352
11.2.2	La construction	353
11.2.3	Un exemple	353
11.2.4	La fin de la construction et l'élagage	354
11.3	Le <i>boosting</i> d'un algorithme d'apprentissage	354
11.3.1	Plusieurs experts valent mieux qu'un	354
11.3.2	Le premier algorithme de <i>boosting</i>	355
11.3.3	Le <i>boosting</i> probabiliste et l'algorithme ADABOOST	357

11.3.4	Les propriétés de l'algorithme ADABOOST	358
11.3.5	L'utilisation du boosting	360
11.3.6	Boosting et théorie <i>PAC</i>	360
11.3.7	Le « bagging »	361
12	L'apprentissage de réseaux bayésiens	363
12.1	Les réseaux d'inférence bayésiens	364
12.1.1	Définitions et notations	366
12.1.2	La d-séparation	366
12.1.3	Définition formelle d'un réseau bayésien	368
12.2	Les inférences dans les réseaux bayésiens	368
12.2.1	Schémas d'inférence	369
12.2.2	La d-séparation généralisée	372
12.3	L'apprentissage des réseaux bayésiens	374
12.3.1	Apprentissage avec structure connue et données complètes	375
12.3.2	Apprentissage avec structure inconnue et données complètes	376
12.3.3	Apprentissage en présence de données incomplètes	379
12.3.4	Apprentissage avec structure connue et données incomplètes	379
12.3.5	Apprentissage avec structure inconnue et données incomplètes	381
13	L'apprentissage de modèles de Markov cachés	385
13.1	Les modèles de Markov observables	388
13.2	Les modèles de Markov cachés (HMM)	389
13.2.1	Définition	389
13.2.2	Pourquoi faut-il des variables cachées?	389
13.2.3	Notations	391
13.2.4	Deux types de HMM	392
13.2.5	Comment un HMM engendre une séquence	393
13.3	Les HMM comme règles de classification de séquences	393
13.3.1	Les trois problèmes des HMM	393
13.3.2	Les HMM et la classification bayésienne	394
13.4	L'évaluation de la probabilité d'observation	395
13.5	Le calcul du chemin optimal: l'algorithme de Viterbi	397
13.6	L'apprentissage	400
13.7	Approfondissements	406
13.8	Applications	407
IV	Apprentissage par approximation et interpolation	409
14	L'apprentissage bayésien et son approximation	411
14.1	L'apprentissage bayésien	413
14.1.1	Présentation	413
14.1.2	Un petit retour en arrière	415
14.1.3	L'apprentissage bayésien d'une règle de classification	415
14.1.4	La classification bayésienne est optimale en moyenne...	416
14.1.5	...mais on ne peut que l'approcher.	416
14.1.6	La règle bayésienne et la régression aux moindres carrés	418
14.1.7	La règle bayésienne et la minimisation de l'entropie croisée	419

14.1.8	La règle bayésienne et la longueur minimale de description	420
14.1.9	L'apprentissage bayésien non supervisé	422
14.2	Les méthodes paramétriques	422
14.2.1	L'estimation par maximum de vraisemblance	422
14.2.2	L'estimation des paramètres d'une distribution gaussienne	423
14.2.3	Des hypothèses simplificatrices	427
14.2.4	Les cas non gaussiens et multigaussiens	428
14.2.5	La prédiction bayésienne de la distribution des paramètres	428
14.3	L'apprentissage bayésien non paramétrique	431
14.3.1	Généralités : le problème de l'estimation locale d'une densité	431
14.3.2	Les fonctions noyau et les fenêtres de Parzen	432
14.3.3	Les k -plus proches voisins (k -ppv)	435
14.4	Les méthodes semi paramétriques	445
14.4.1	La discrimination logistique	445
14.4.2	Les mélanges de distributions	447
14.4.3	Le cas des réseaux connexionnistes et des arbres de décision	448
15	La classification non supervisée et la découverte automatique	451
15.1	La classification hiérarchique de données numériques	453
15.1.1	Généralités	453
15.1.2	Un algorithme général de classification hiérarchique	456
15.1.3	L'indice du lien simple	456
15.1.4	L'indice de la distance entre centres de gravité	457
15.1.5	L'indice de Ward	457
15.1.6	L'indice de la vraisemblance du lien	458
15.1.7	Le choix du nombre de classes	458
15.2	La classification non hiérarchique de données numériques	458
15.2.1	La méthode des k -moyennes	458
15.2.2	L'estimation d'une somme pondérée de distributions gaussiennes	460
15.2.3	Un exemple	460
15.3	La classification de données symboliques	462
15.3.1	Les données binaires et catégorielles	462
15.3.2	Les attributs nominaux : la représentation attribut-valeur	463
15.3.3	Les données logiques	464
15.4	La découverte automatique	464
15.4.1	Présentation	464
15.4.2	La découverte de fonctions simples	465
15.4.3	Découverte de lois plus complexes	466
15.4.4	Traitement des données bruitées	467
15.4.5	Découverte de lois avec plus de deux variables	468
15.4.6	Améliorations ultérieures	470
15.5	La découverte non supervisée d'associations complexes d'attributs	470
15.5.1	Les associations d'attributs binaires : définitions	470
15.5.2	L'apprentissage des associations	471
15.5.3	Découverte de suites temporelles dans les données	473
15.6	Le coapprentissage et les mélanges d'exemples supervisés et non supervisés	479
15.6.1	Le cas de deux jeux indépendants d'attributs : le coapprentissage	479

15.6.2	L'utilisation de l'algorithme <i>EM</i>	480
16	L'apprentissage de réflexes par renforcement	483
16.1	Description du problème	485
16.1.1	La modélisation d'un agent en action dans le monde	485
16.1.2	Les notions fondamentales	487
16.1.3	Les problèmes et les grandes approches	490
16.2	Si tout est connu : l'utilité de la fonction d'utilité	491
16.3	L'apprentissage des fonctions d'utilité quand l'environnement est connu	492
16.3.1	L'évaluation d'une politique par propagation locale d'information	493
16.3.2	Un théorème conduisant à l'amélioration de politique	494
16.3.3	Processus itératif d'amélioration de politique	495
16.4	Si rien n'est connu : la méthode de Monte-Carlo	496
16.5	Le meilleur des deux mondes : la méthode des différences temporelles	497
16.5.1	L'évaluation suivant la méthode des différences temporelles	497
16.5.2	L'amélioration de politique avec les différences temporelles	498
16.5.3	SARSA : Une méthode d'amélioration « sur politique »	499
16.5.4	Le <i>Q-learning</i> : Une méthode d'amélioration « hors politique »	499
16.5.5	TD(λ) : les méthodes de différences temporelles à plusieurs pas	500
16.6	La généralisation dans l'apprentissage par renforcement	501
16.6.1	Le problème	501
16.6.2	Généralisation par approximation de la fonction d'utilité	502
16.6.3	Méthodes de généralisation par partition de l'espace	504
16.6.4	Méthodes directes d'apprentissage de politique	506
16.7	Le cas des environnements partiellement observables	506
16.8	Exemples d'application	507
16.8.1	Le TD-Gammon	507
16.8.2	Applications au contrôle et à la robotique	508
16.9	Bilan et perspectives	508
V	Approfondissements et annexes techniques	511
17	Approfondissement sur l'analyse de l'induction	513
17.1	L'analyse de l'induction de Vapnik	513
17.1.1	Cas où $ \mathcal{H} = \infty$ et $\mathcal{F} \subseteq \mathcal{H}$	514
17.1.2	Fonction de croissance et dimension de Vapnik-Chervonenkis	515
17.1.3	Le lemme de Sauer : un lemme sauveur	517
17.1.4	L'analyse de Vapnik et Chervonenkis pour des fonctions quelconques	520
17.1.5	Discussion	523
17.2	Les principes inductifs avec contrôle de l'espace des hypothèses	524
17.2.1	La minimisation du risque structurel : <i>SRM</i>	524
17.2.2	La théorie de la régularisation	525
17.2.3	La théorie de l'estimation bayésienne	528
17.3	L'induction par compression d'information	529
17.3.1	Un exemple	530
17.3.2	La théorie de l'induction selon Solomonoff	530

17.3.3	La complexité de Kolmogorov	531
17.3.4	Le principe de longueur de description minimale (<i>MDLP</i>)	532
17.3.5	Analyse: compression et pouvoir inductif	534
17.4	L'induction en débat	536
17.4.1	Le <i>no-free-lunch theorem</i> : toutes les méthodes se valent!?	536
17.4.2	Le <i>no-free-lunch theorem</i> et l'analyse de Vapnik: une contradiction?	541
17.5	Discussion sur l'analyse classique. Variantes et perspectives	542
17.5.1	D'autres modèles d'apprentissage	544
17.5.2	D'autres types d'analyses	546
18	Annexes techniques	551
18.1	Exemples de fonctions de perte en induction	551
18.1.1	La reconnaissance de formes ou classification	551
18.1.2	La régression	552
18.1.3	L'estimation de densité	553
18.2	Optimisation par descente de gradient	554
18.3	La rétropropagation du gradient de l'erreur	557
18.3.1	Notations	557
18.3.2	Fonctionnement du système	557
18.3.3	Calcul du gradient	558
18.4	Estimation d'une densité de probabilité en un point	560
18.5	L'estimation des paramètres d'une distribution gaussienne	561
18.6	Pourquoi et comment la règle du PPV converge-t-elle?	562
18.6.1	Pourquoi?	562
18.6.2	Comment?	562
18.7	Le calcul de l'intervalle de confiance de l'estimation de la probabilité d'une règle de classification	563
18.8	Pourquoi la règle de décision bayésienne est-elle optimale?	564
18.9	Apprentissage par estimation-maximisation.	565
18.9.1	Un exemple	565
18.9.2	Application de l'algorithme <i>EM</i> à l'exemple	565
18.9.3	Statistique suffisante	566
18.9.4	Plus généralement	566
18.9.5	Retour sur l'exemple	566
18.9.6	L'apprentissage des paramètres des HMM	568
18.9.7	L'apprentissage des paramètres de distributions multigaussiennes	568
	Bibliographie	571
	Index	586